УДК 004.622
С. И. Пухов
ФГАОУ ВО «Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина», г. Екатеринбург, Россия

СОЗДАНИЕ СИСТЕМЫ СБОРА И ОБРАБОТКИ ОТКРЫТЫХ ДАННЫХ С РЕСУРСОВ СЕТИ ИНТЕРНЕТ

Аннотация

Данная работа посвящена созданию системы автоматического сбора и обработки открытых данных с ресурсов сети интернет и несет в себе практическую значимость в задачах анализа текста. Во введении обосновывается актуальность выбранной темы, формулируются цель и задачи исследования, указывается объект и предмет исследования. Рассматривается такая задача, как сбор и первичная обработка текстовых данных с последующим анализом. Сбор данных является первоочередной задачей, так как открытые данные с ресурсов сети интернет не структурированы и нуждаются в обработке. Автор предоставляет систему обработки HTML страниц и файлов с ресурсов образовательных учреждений, а также приводит практическое применение данного подхода на реальных данных открытых ресурсов с помощью созданной системы. Система поможет структурировать открытые данные с ресурсов сети интернет, а также провести анализ собранных данных.

Ключевые слова: *сбор данных, обработка данных, обрабока html-страниц, образовательные учреждения, министерство образования, анализ данных.*

Abstract

This work is devoted to the creation of a system for automatic collection and processing of open data from Internet resources and bears practical significance in problems of text analysis. In the introduction, the relevance of the selected topic is substantiated, the goal and tasks of the research are formulated, the object and subject of the research are indicated. We consider such a task as the collection and initial processing of text data with subsequent analysis. Data collection is a priority, since open data from Internet resources are not structured and need to be processed. The author provides a system for processing HTML pages and files from the resources of educational institutions, and also leads the practical application of this approach to real data of open resources with the help of the created system. The system will help to structure the open data from the Internet resources, as well as analyze the collected data.

Key words: *data collection, data processing, processing of html-pages, educational institutions, the Ministry of Education, data analisis.*

Введение. Все данные, представленные в глобальной сети Интернет, можно назвать неструктурированными, ввиду индивидуальности и специфичности архитектуры каждого ресурса в отдельности. В основном, такие данные – это HTML страницы, в которых хранится текстовая информация, а также ссылки на определенные файлы, хранящиеся на сервере. В настоящее время, в связи с постоянным ростом информации во всемирной паутине, необходимо развивать технологии, которые позволят собирать и обрабатывать информацию автоматически. Если исследования ведутся с большой выборкой, то для того, чтобы накопить достаточное количество материала, могут уйти недели или месяцы кропотливого труда. Данная работа посвящена разработке системы автоматического

сбора и обработки информации из открытых источников сети интернет. Эта система позволит собирать статистику из образовательных учреждений с построением зависимостей, графиков и диаграмм, а также с возможностью прогнозирования дальнейших результатов каждого заведения.

Целью данной статьи является разработка системы по сбору и обработки данных о самообследовании с ресурсов образовательных учреждений.

Для достижения поставленной цели потребовалось решить следующие задачи:

1) разработка базы данных для хранения данных с ресурсов;
2) разработка приложения по сбору данных;
3) разработка системы анализа данных.

Данные, которые требовалось собрать и обработать, представляют собой набор параметров из отчётов о результатах самообследования в различных образовательных учреждениях. Объектом данных является отчет о самообследовании, который формируется ежегодно в образовательных учреждениях. Таким образом, данных лежат на серверах учреждениях в файлах формата pdf, doc или xls. Источником данных являются сайты образовательных учреждений.

Описание разработки. Для создания системы требуются знания таких языков как HTML, C# и SQL, умение проводить синтаксический анализ (парсинг) файлов xls, doc и pdf.

Процесс разработки. На первом этапе разработки системы для сбора и обработки информации с образовательных учреждений были определены функциональные возможности, которые необходимы для достижения поставленных целей и задач разработки. Разрабатываемая система должная обеспечивать следующие возможности:

– вход на сервер образовательного учреждения;
– поиск файла самообследования в формате pdf, doc или xls на сервере учреждений (рисунки 1 и 2);
– скачивание файлов самообследования;
– синтаксический анализ скачиваемого файла формате pdf, doc или xls с последующим заполнением базы данных;
– проведение анализа информация из таблиц базы данных.

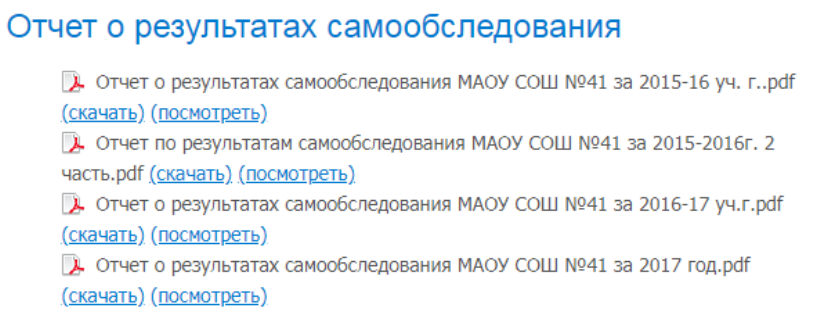

Рис. 1. Примерный вид файлов формата pdf на сайте образовательного учреждения

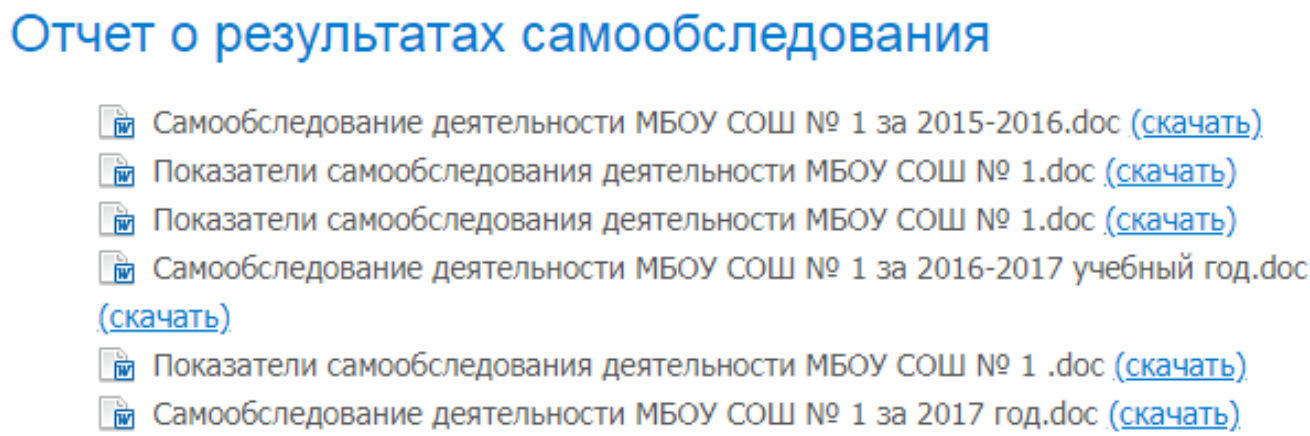

Рис. 2. Примерный вид файлов формата doc на сайте образовательного учреждения

Основным требованием, которым должна отвечать разрабатываемая система – это возможность получения краткого отчета о результатах того или иного образовательного учреждения по конкретному шаблону и заполнение базы данных нужными параметрами.

Дальнейшим шагом разработки является интеллектуальный анализ собранных данных.

Интеллектуальный анализ данных представляет собой процесс обнаружения пригодных к использованию сведений в крупных наборах данных. В интеллектуальном анализе данных применяется математический анализ для выявления закономерностей и тенденций, существующих в данных [1]. Обычно такие закономерности нельзя обнаружить при традиционном просмотре данных, поскольку связи слишком сложны, или из-за чрезмерного объема данных.

Для интеллектуального анализа существует ряд алгоритмов, наиболее распространенные:

– *алгоритмы классификации* осуществляют прогнозирование одной или нескольких дискретных переменных на основе других атрибутов в наборе данных.

– *регрессивные алгоритмы* осуществляют прогнозирование одной или нескольких непрерывных переменных, например, прибыли или убытков, на основе других атрибутов в наборе данных.

– *алгоритмы сегментации* делят данные на группы или кластеры элементов [2], имеющих схожие свойства.

– *алгоритмы взаимосвязей* осуществляют поиск корреляции между различными атрибутами в наборе данных. Наиболее частым применением этого типа алгоритма является создание правил взаимосвязи, которые могут использоваться для анализа потребительской корзины.

– *алгоритмы анализа последовательностей* обобщают часто встречающиеся последовательности в данных, например, поток данных в Интернете.

С помощью алгоритмов, представленных выше, был произведен интеллектуальный анализ данных, полученных с разных образовательных учреждений. Пример кластерного анализа продемонстрирован на рисунке 3.

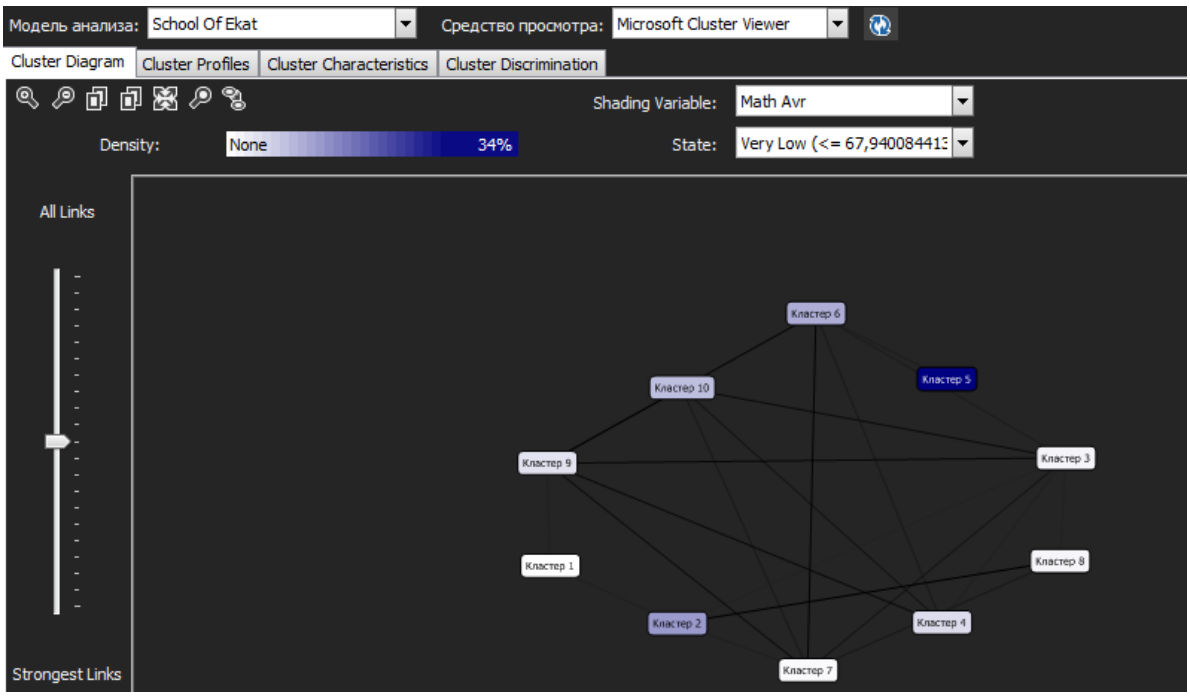

Рис. 3. Кластерный анализ данных

Таким образом были проведены несколько разных типов анализов для наших различных целей.

Заключение. Разработана система, обеспечивающая поиск отчетов самообследования, по ключевым словам, на серверах образовательных учреждений, с последующей обработкой полученных файлов и их глубоким анализом.

Список использованных источников

1. Основные понятия интеллектуального анализа данных. [Электронный ресурс] / 2017 – Режим доступа: https://docs.microsoft.com/ru-ru/sql/analysis-services/data-mining/data-mining-algorithms-analysis-services-data-mining?view=sql-analysis-services-2017.

2. Алгоритмы интеллектуального анализа данных (службы Analysis Services – интеллектуальный анализ данных). [Электронный ресурс] – Режим доступа: https://msdn.microsoft.com/ru-ru/library/ms175595(v=sql.120).aspx.